8.4 An 8.4ns Column-Access 1.3Gb/s/pin DDR3 SDRAM with an 8:4 Multiplexed Data-Transfer Scheme

Hiroki Fujisawa[1], Shuichi Kubouchi[2], Koji Kuroki[1], Naohisa Nishioka[1], Yoshiro Riho[1], Hiromasa Noda[1], Isamu Fujii[2], Hideyuki Yoko[2], Ryuuji Takishita[2], Takahiro Ito[2], Hitoshi Tanaka[2], Masayuki Nakamura[1]

[1]ELPIDA Memory, Sagamihara, Japan
[2]Hitachi ULSI Systems, Kokubunji, Japan

The column-access time of a 512Mb DDR3 SDRAM, implemented in a 90nm dual-gate CMOS process, is reduced by 2.9ns; from 11.3 to 8.4ns. This is achieved through an 8:4 multiplexed data-transfer scheme in a quasi-4b prefetch operation. A dual-clock latency counter reduces the cycle time by 30%, from 1.7 to 1.2ns. A multiple-ODT merged output driver enables 3% accuracy through ZQ self-calibration at any R_{on} and R_{tt} values. Using these techniques, a 1.3Gb/s/pin operation with a 1.36V supply and a column latency of 6, is achieved.

For data rates of over 1.3Gb/s/pin, the prefetch size and the column latency (CL) of the DRAM have steadily increased from DDR1 to DDR3 [1, 2]. However, the large prefetch size and the long latency strongly affect the memory system performance. Therefore, a higher DRAM core speed that can be achieved for example by reducing the column-access time (t_{AA}), is needed. To respond to this market demand as flexibly as possible, the proposed approach is to reduce t_{AA} to <9ns and enable quasi-4b-prefech data transfer in DDR3. The block and timing diagrams of a proposed 8:4 multiplexed data-transfer scheme that enables high-speed and high-frequency data transfer over global-I/O (GIO) lines are shown in Figs. 8.4.1 and 8.4.2, respectively. The GIO lines are quite long, up to 9000µm, and the amount of simultaneous data transfer in the 8b prefetch is twice that of a 4b prefetch operation. Hence, wide-pitch and fully shielded GIO lines are needed. In this scheme, by adding a 4b latch in the main amplifier and reducing the number of GIO lines by half, a 2.6µm-pitch and shielded GIO lines are achieved without any area penalty. The first half of the 8b prefetch data is transferred immediately to the GIO lines using the ¢0 signal and prefetch address. The latter half of the 8b prefetch data is then transferred from the 4b latch to the GIO lines by ¢1 signals with a two-clock delay. Therefore, the peak current of the GIO output buffers is reduced by half. Furthermore, when the burst length is 4, the peak current is almost the same as that of a burst length of 8, and this reduces the timing design complexity (such as the GIO data latch margin in a quasi-4b prefetch operation). As Fig. 8.4.4a shows, t_{AA} is reduced by 2.9ns to 8.4ns, which is low enough for stable 1.3Gb/s/pin operation with a column latency of 6. This reduction consisted of a 1.0ns reduction through the 8:4 multiplexed data-transfer scheme and a 1.9ns reduction through the dual-gate CMOS technology [3].

As the operating frequency increases, so does the latency number. For example, the additive latency (AL) is seven values (0 and 4 to 9) in DDR3 operation, which is more than twice that of DDR2 operation. Therefore, the decrease in the internal timing margin and the increased power consumption of the latency counter become serious problems. In a conventional latency counter, a serial flip-flop and a selector are used to count latency, and the frequencies of the external clock and internal clock for the latency counter are the same [1]. The input command signal of a 10b latency counter must be transferred via a 7-input AL selector in one clock period, so the minimum cycle time must be no more than 1.7ns to ensure the latch margin and external-clock jitter margin are adequate. To overcome these problems, the dual-clock latency counter circuits shown in Fig. 8.4.3 are developed. Two sets of 5b latency counters and four 3-input selectors controlled by dual-phase 1-shot clock signals are connected in parallel. An external command is input into either of the one-sided flip-flops and transferred to a latency counter. In the case of even-number latency, only one side of the 5b counter is used and all of the transfer time between flip-flops is equal to two clocks, which is almost twice that of the conventional scheme. In the case of odd-number latency, the command signal must be transferred from the counter set on one side to the other side via a 3-input selector. To prevent the transfer time from becoming one clock, D-latch circuits controlled by an inverted dual-phase clock are added. Therefore, the transfer time from node A to node C is equal to three clocks and that is almost 1.5 times that of the conventional scheme. Thus, the minimum cycle time is 1.2ns at 1.35V, which is 0.5ns shorter than that of the conventional scheme. There is enough margin for 1.3Gb/s/pin and a potential for 1.6Gb/s/pin DDR3 operation (Fig. 8.4.4b). Furthermore, the power consumption of this latency counter is almost half of that of the conventional scheme. Because the total number of flip-flops is almost the same, but the flip-flop operating frequency is halved by using the dual-phase clock.

ODT is widely used, because it improves the signal integrity in high-frequency systems such as DDR2 or DDR3, but the input capacitance of the DQ pin (Cio) is increased by adding an extra ODT driver. Figure 8.4.5a shows a partially ODT-merged output driver [1]. This circuit is one of the best solutions to minimize Cio, because using both ZQ self-calibration and off-chip driver (OCD) calibration combines the good linear ODT driver and the less linear output driver. However, this circuit makes the calibration process more complex and the OCD calibration accuracy depends on the memory system. To overcome these problems, the multiple-ODT merged output driver, where the ODT buffer and output driver are fully merged, is proposed. Figure 8.4.5b shows the block diagram of this circuit, where the output driver is composed of the same 240Ω units. The value of output driver impedance (R_{on}) and ODT effective resistance (R_{tt}) are controlled by changing the number of 240Ω units that are simultaneously activated. Therefore, multiple R_{tt} values (20, 30, 40, 60, and 120Ω) are possible without increasing Cio, and the R_{tt} value can be flexibly chosen for every DRAM system composition. The calibrations of R_{on} and R_{tt} are done simultaneously through ZQ self-calibration. The measured characteristics of the multiple-ODT merged output driver after ZQ self-calibration are shown in Fig. 8.4.5c. The fluctuation of R_{on} and R_{tt} is no more than 3%, where a target accuracy of ±10% can be achieved at any R_{on} and R_{tt} values. The Cio of the proposed circuits is 2.25pF, almost 30% less than that of the separated ODT and output driver.

Figure 8.4.7 shows a micrograph of a chip fabricated in a 90nm 3AL 1W dual-gate CMOS process. All the JEDEC standard DDR3 functions are verified. The maximum data rate at 1.36V is 1.3Gb/s/pin (Fig. 8.4.6), which is enough for DDR3-1333 operation with column latency of 6.

References:

[1] C. Park, et al., "A 512Mbit, 1.6Gbps/pin DDR3 SDRAM Prototype with Cio Minimization and Self-Calibration Techniques," *Dig. Symp. VLSI Circuits*, pp. 370-373, June, 2005.

[2] H. Fujisawa, et al., "1.8-V 800-Mb/s/pin DDR2 and 2.5-V 400-Mb/s/pin DDR1 Compatibly Designed 1-Gb SDRAM," IEEE *J. Solid-State Circuits*, vol. 40, pp. 862-869, Apr., 2005.

[3] K. Saino, et al., "A Novel W/WNx/Dual-gate CMOS Technology for High-speed DRAM Having Enhanced Retention Time and Reliability," *IEDM Dig. Tech. Papers*, pp. 415-418, Dec., 2003.

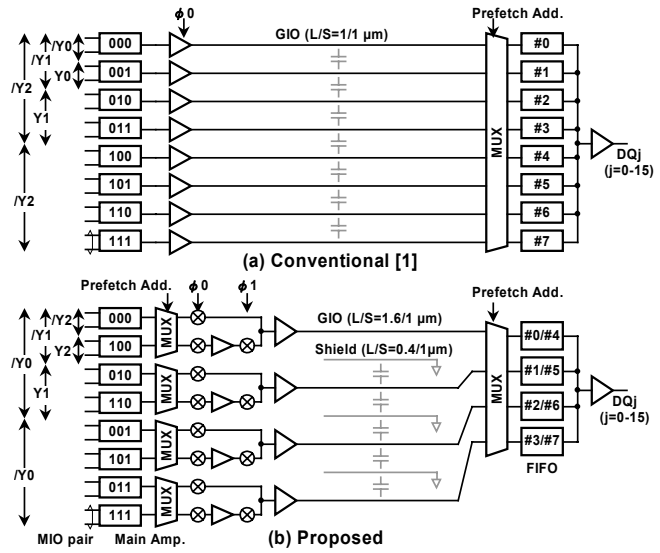

Figure 8.4.1: Block diagram of 8:4 multiplexed data-transfer scheme.

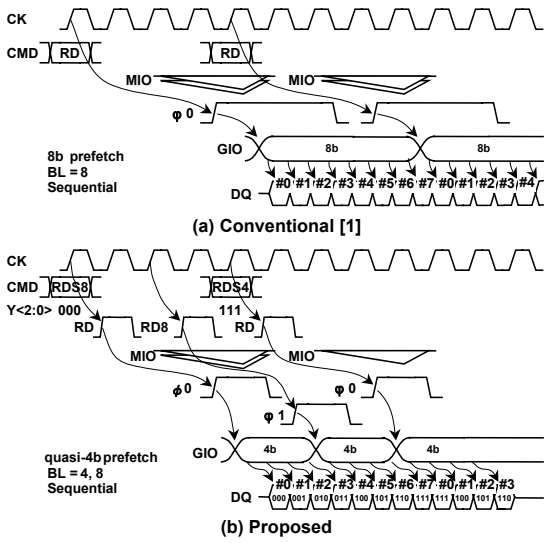

Figure 8.4.2: Timing diagram of 8:4 multiplexed data-transfer scheme.

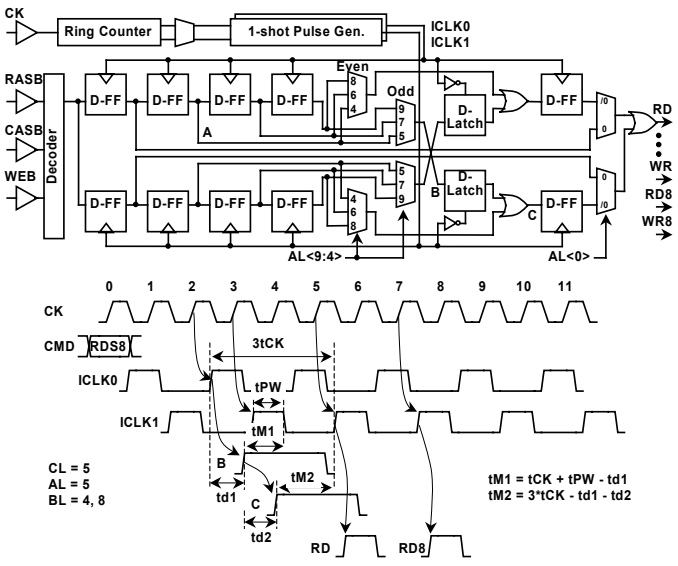

Figure 8.4.3: Dual clock latency counter.

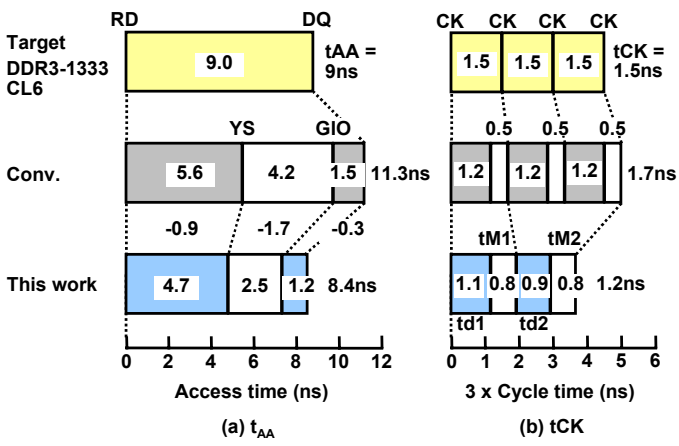

Figure 8.4.4: Performance improvement.

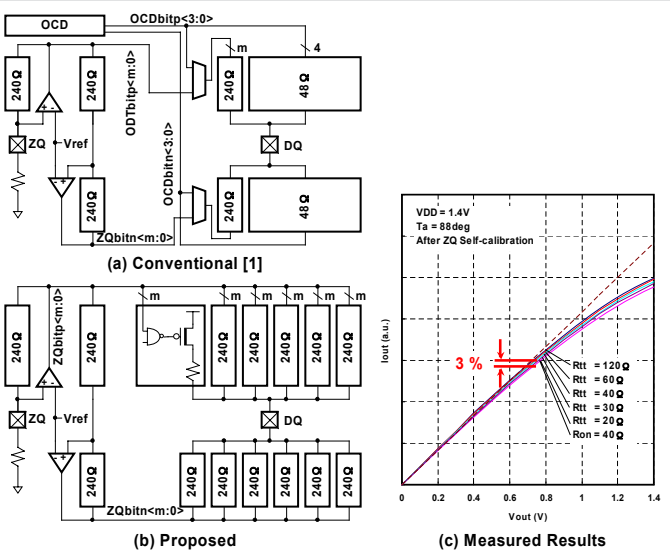

Figure 8.4.5: Multiple-ODT-merged output driver.

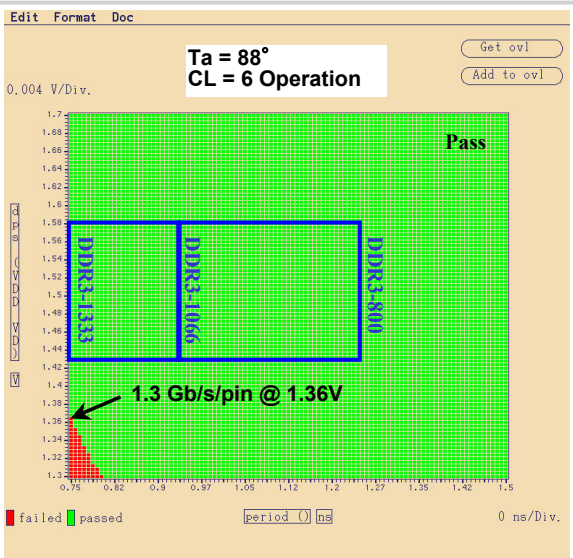

Figure 8.4.6: Shmoo plot.

Continued on Page 645

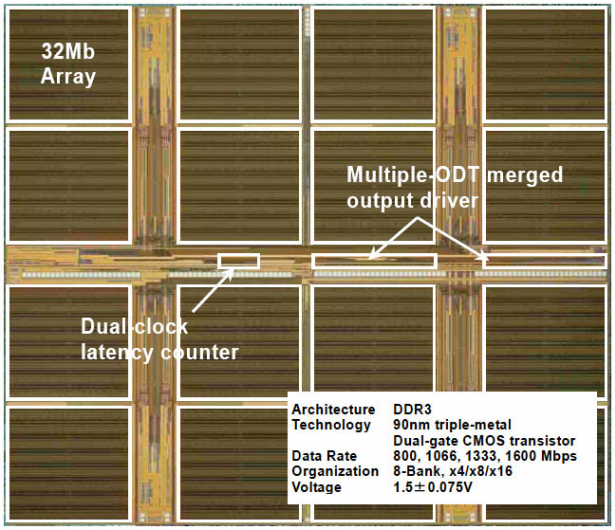

Figure 8.4.7: Chip micrograph.

